The Capacity of Linear Channels with Additive Gaussian Noise

By R. K. MUELLER and G. J. FOSCHINI

(Manuscript received August 8, 1969)

The standard method of computing the mutual information between two stochastic processes with finite energy replaces the processes with their Fourier coefficients. This procedure is mathematically justified here for random signals $w_t(\omega)$ square-integrable in the product space $t \times \omega$ where $t \in [0, T]$ and ω is an element of a probability space. A natural notion of the sigma field generated by $w_t(\omega)$ is presented and it is shown to coincide with the sigma field generated by the random Fourier coefficients of $w_t(\omega)$ in any complete orthonormal system in $L_2[0, T]$. This justifies the use of Fourier coefficients in mutual information computations.

Capacity is calculated for finite and infinite-dimensional channels, where the output signal consists of a filter (general Hilbert–Schmidt operator) operating on the input signal with additive Gaussian noise. The finite-dimensional optimal signal is obtained. In the infinite-dimensional case capacity can be approached arbitrarily closely with finite-dimensional inputs. The question of the existence of an infinite-dimensional signal which achieves capacity is considered. There are channels for which no signal achieves capacity. Some results are obtained when the noise coordinates are independent in the eigensystem of the filter.

I. INTRODUCTION

In this paper, we attack a general form of the classical problem of determining the capacity of a linear channel with additive noise. Structurally we have

$$r_t(\omega) = \int_0^T \mathbf{G}(t, \tau) s_\tau(\omega)\, d\tau + n_t(\omega) \tag{1}$$

where the random signals, noise $[n_t(\omega)]$, input $[s_t(\omega)]$, and output $[r_t(\omega)]$ are all defined on $0 \leqq t \leqq T$. All signals as well as the kernel of the channel operator are assumed square integrable in the appropriate

product spaces. The noise process, the channel operator, and an average power restriction on $s_t(\omega)$ are assumed to be given. In Section III we begin by defining the capacity of a channel. Our definition is motivated by, but is not a special case of, the generalization of Sbannon's notion of capacity that has been indicated by Kolmogorov. The argument for the naturalness of our definition is that any of the above processes can be replaced by their random Fourier coefficients from any expansion using complete orthonormal functions in $L_2[0, T]$. We solve the above problem when $n_t(\omega)$ is Gaussian and independent of $s_t(\omega)$. In Section IV we show that for finite-dimensional inputs there always exists an $s_t(\omega)$ for which capacity is achieved and we find it. The infinite-dimensional case is solved in Section V as a limit of finite-dimensional cases.

II. FUNDAMENTALS

Fundamental to the notion of capacity is the notion of mutual information. We begin with Kolmogorov's definition of the mutual information of two event σ-fields contained in a universal σ-field. Let $\mathcal{A}$ and $\mathcal{B}$ denote two sub σ-fields of a σ-field S_Ω in a probability space (Ω, S_Ω, P). Let α and β denote arbitrary partitions of Ω into a finite number of $\mathcal{A}$ and $\mathcal{B}$ measurable sets A and B. The mutual information $I(\mathcal{A}, \mathcal{B})$ of $\mathcal{A}$ and $\mathcal{B}$ is

$$I(\mathcal{A}, \mathcal{B}) = \sup_{\alpha,\beta} \sum_{A \varepsilon \alpha} \sum_{B \varepsilon \beta} P(A \cap B) \log_e \frac{P(A \cap B)}{P(A)P(B)}. \tag{2}$$

We define $0 \log 0 = 0$. This sum does not decrease as α and β are refined. It can be shown that $I(\mathcal{A}, \mathcal{B}) \geqq 0$ with equality if and only if $\mathcal{A}$ and $\mathcal{B}$ are independent. The nonnegativity and other important properties of I are presented in Ref. 1.

Let $\mathfrak{X}$ be a measurable space with σ-field denoted by $\mathfrak{D}$. A function $\zeta(\omega)$ from (Ω, S_Ω, P) to $\mathfrak{X}$ for which each $D \varepsilon \mathfrak{D}$ has a preimage in S_Ω is called a measurable function.

Let T be an arbitrary index set and let E^1 denote the real line. Endow $\Pi_{t \varepsilon T} E^1$ with the product topology and consider its measurable sets to be the smallest σ-field containing the topology. We are interested in measurable functions from Ω to $\Pi_{t \varepsilon T} E^1$. For our purposes T is either countable or a real compact interval.

Suppose ξ and η are measurable functions from Ω to $\Pi_{t \varepsilon T} E^1$. Then by the mutual information of ξ and η, $I(\xi, \eta)$, we mean the mutual information between tbe smallest σ-fields with respect to which ξ and η are measurable. We denote these respective σ-fields by $\mathcal{A}_\xi$ and $\mathcal{A}_\eta$.

Let $\zeta(\omega)$ denote any measurable function from Ω to $\Pi_{t\epsilon T} E^1$. We define the probability distribution P_ζ of $\zeta(\omega)$. The domain of P_ζ is the measurable sets in $\Pi_{t\epsilon T} E^1$. Let Q be such a measurable set. Then

$$P_\zeta(Q) = P\{\omega : \zeta(\omega) \in Q\}. \tag{3}$$

If ξ and η are each measurable functions from Ω to $\Pi_{t\epsilon T_1} E^1$ and $\Pi_{t\epsilon T_2} E^1$ respectively then (ξ, η) is a measurable function from Ω to $\Pi_{t\epsilon T_1} E^1 \times \Pi_{t\epsilon T_2} E^1$ and its distribution function is denoted $P_{\xi,\eta}$. It is called the joint distribution of ξ and η. We can now give an alternate definition of mutual information between ξ and η. Let $\gamma(\delta)$ denote arbitrary partitions of $\Pi_{t\epsilon T_1} E^1(\Pi_{t\epsilon T_2} E^1)$ into a finite number of measurable sets $C(D)$. The mutual information $I(\xi, \eta)$ is

$$I(\xi, \eta) = \sup_{\gamma, \delta} \sum_{C \epsilon \gamma} \sum_{D \epsilon \delta} P_{\xi\eta}(C \times D) \log \frac{P_{\xi\eta}(C \times D)}{P_\xi(C) P_\eta(D)}. \tag{4}$$

Recall that the inverse image under a measurable function of a σ-field is a σ-field. So it becomes apparent that the two definitions for $I(\xi, \eta)$ are equivalent.

We review without proof some fundamental propositions that will be of use to us later. The following is a result of work by I. M. Gelfand, A. M. Yaglom, and A. Perez.

Theorem: If $P_{\xi\eta}$ is not absolutely continuous with respect to the product measure $P_\xi \times P_\eta$ then $I(\xi, \eta) = \infty$. If $P_{\xi\eta}$ is absolutely continuous with respect to $P_\xi \times P_\eta$, then letting $dP_{\xi\eta}/d(P_\xi \times P_\eta)$ denote the Radon–Nikodym derivative of $P_{\xi\eta}$ with respect to $P_\xi \times P_\eta$ we have

$$I(\xi, \eta) = \int_\Omega \left[\log \frac{dP_{\xi\eta}}{d(P_\xi P_\eta)} \right] dP_{\xi\eta} . \tag{5}$$

Proof: See Ref. 2.

Theorem: Let A be a linear transformation in a k-dimensional vector space and let ξ be a k-dimensional random vector. Then

$$I(\xi, \eta) \geqq I(A\xi, \eta) \tag{6}$$

holds for any random vector η, with equality if the transformation A is nonsingular.

Proof: See Ref. 3.

Theorem: If $I(\xi, \xi) < \infty$, then P_ξ is purely atomic.

Proof: See Ref. 4.

Theorem: *If* $\xi = (\xi_1, \xi_2, \cdots)$, *then*

$$I(\xi, \eta) = \lim_{n\to\infty} I[(\xi_1, \cdots, \xi_n), \eta]. \tag{7}$$

Proof: See Ref. 4.

III. MUTUAL INFORMATION BETWEEN TWO PROCESSES IN $L_2\{(\Omega, S_\Omega, P) \times ([0, T], L, m)\}$

Let $\xi_t(\omega)$ be square integrable on $t \times \omega$. We term $\xi_t(\omega)$ a stochastic process. Notice that it differs from the standard definition of a stochastic process in two ways. First, it is an equivalence class of equal almost everywhere functions in $(t \times \omega)$. Second, not all functions in the equivalence class are stochastic processes in the sense of Ref. 5.; that is, for each t we do not have a random variable but only for almost all t. We assume $E\{\xi_t(\omega)\} = 0$. By Schwarz's inequality and Fubini's theorem it follows that $E\{\xi_{t_1}(\omega)\xi_{t_2}(\omega)\} \in L_2(t \times t)$. If $\eta_t(\omega)$ and $\zeta_t(\omega)$ are processes of the same type as $\xi_t(\omega)$, $I[\eta_t(\omega), \zeta_t(\omega)]$ is not well defined since $\mathcal{A}_\eta$ and $\mathcal{A}_\zeta$ are not well defined. Because of the central role of these processes in modeling random signals with a finite average power we make $\mathcal{A}_\eta$ and $\mathcal{A}_\zeta$ and hence $I[\eta_t(\omega), \zeta_t(\omega)]$ meaningful here. We need to appeal to the following:

Theorem (F. Riesz): *Let* f_n *converge in measure to* f. *Then there exists a subsequence* f_{n_k} *converging to* f *almost everywhere.*

Proof: See Ref. 6.

Suppose f_n converges in mean square to f. Since convergence in mean square implies convergence in measure, the limit of the subsequence guaranteed by Riesz's theorem is f in the sense that the limit and f agree almost everywhere. This last comment is important since Kolmogorov has given examples of functions g which possess an orthogonal expansion g_n converging in mean square to g, yet pointwise almost everywhere convergence does not occur.

Unless stated otherwise all σ-fields mentioned in the remainder of this section are assumed to be completed. The following new definition is the key to making $\xi_t(\omega)$ meaningful in the information theory sense.

Definition: By the σ-field $\mathcal{A}$ generated by $\xi_t(\omega)$ we mean the smallest σ-field $\mathcal{A}$ satisfying $\xi_t(\omega)$ is $(\Omega, \mathcal{A}) \times ([0, T], L)$ measurable, where L is the sigma field of Lebesgue measurable subsets of $[0, T]$. (This statement is definitive since $\xi_t(\omega)$ is $(\Omega, S_\Omega) \times ([0, T], L)$ measurable and the intersection of σ-fields is a σ-field.)

Proposition I: *Suppose*

$$\sum_{i \geq 1} a_i(\omega)\phi_i(t) = \xi_t(\omega) \tag{8}$$

in the mean square sense in the product space (where $a_i(\omega) = \int \xi_t(\omega)\phi_i(t)\, dt$ *and* $\phi_i(t)$ *are orthonormal on* $[0, T]$*). If* $a(\omega) = [a_1(\omega), a_2(\omega), \cdots]$ *then* $\mathfrak{A}_\xi = \mathfrak{A}_a$.

Proof: Since the expansion converges to $\xi_t(\omega)$ in mean square in the product space, it converges in measure. By F. Riesz's theorem we can find $n_1 < n_2 < \cdots$ so that

$$\lim_{n_k \to \infty} [a_1(\omega)\phi_1(t) + \cdots + a_{n_k}\phi_{n_k}(t)] = \xi_t(\omega). \tag{9}$$

The sum and product of measurable functions is measurable so that each partial sum is $(\Omega, \mathfrak{A}_a) \times ([0, T], L)$ measurable. The limit of measurable functions is measurable so $\xi_t(\omega)$ is $(\Omega, \mathfrak{A}_a) \times ([0, T], L)$ measurable. Thus $\mathfrak{A}_\xi \subset \mathfrak{A}_a$.

Next we project $\xi_t(\omega)$ on $\phi_i(t)$ to get

$$a_i(\omega) = \int \xi_t(\omega)\phi_i(t)\, dt. \tag{10}$$

By Fubini's theorem $a_i(\omega)$ is measurable with respect to every σ-field $\mathfrak{B}$ for which $\xi_t(\omega)$ is $(\Omega, \mathfrak{B}) \times ([0, T], L)$ measurable. But this is true for each i, so $\mathfrak{A}_a \subset \mathfrak{A}_\xi$.

Proposition I is of paramount importance. In the sequel it enables us to replace $\xi_t(\omega)$ by $a(\omega)$ when computing mutual information.

It would seem appropriate to express $\mathfrak{A}_\xi$ without reference to an expansion. The following proposition accomplishes this. However, our proof does resort to an expansion of $\xi_t(\omega)$. Because the proof is similar to the proof of proposition I, we omit it.

Proposition II: *Let* $\{\xi_t^\alpha(\omega)\}$ *denote the class of functions in* $\xi_t(\omega)$. *Then* $\mathfrak{A}_\xi$ *is the smallest* σ*-field containing* $\bigcap_\alpha \mathfrak{A}_{\xi^\alpha}$ [Here we have the only appearance of possibly noncomplete σ- fields (the $\mathfrak{A}_{\xi^\alpha}$)].

We can now define capacity of our noisy linear channel. Let S denote a finite average power restriction on $s_t(\omega)$. Then the capacity of the channel is defined as the supremum of $I[s_t(\omega), r_t(\omega)]$ where the supremum is over all $s_t(\omega)$ satisfying

$$E\left[\frac{1}{T}\int_0^T s_t^2(\omega)\, dt\right] \leq S. \tag{11}$$

We say $\xi_t(\omega)$ is Gaussian if the linear functionals

$$\int_0^T \xi_t(\omega)\phi(t)\,dt \qquad \{\phi(t) \in L_2[0, T]\}$$

are all Gaussian random variables.

IV. THE FINITE-DIMENSIONAL CASE

For a random variable η possessing density p_η the quantity

$$h(\eta) = -\int p_\eta \log p_\eta \tag{12}$$

arises often in mutual information studies. It is called the differential entropy of η.

The following theorem is proved in (Ref. 7).

Theorem: Let p_u be the density of a k-dimensional random variable $\mathbf{u}$. *To maximize*

$$h(\mathbf{u}) = -\int p_u \log p_u$$

subject to the conditions that the mean and dispersion matrix have given values $\boldsymbol{\mu}$ and Γ, choose the normal density

$$Q(\mathbf{u}) = 2\pi^{-k/2} \mid \Gamma \mid^{-\frac{1}{2}} \exp\left[\tfrac{1}{2}(\mathbf{u} - \boldsymbol{\mu})'\Gamma^{-1}(\mathbf{u} - \boldsymbol{\mu})\right],$$

which satisfies the conditions.

We prove a corollary necessary for the sequel.

Corollary: Let p_u be the density of a k-dimensional variable $\mathbf{u}$. *We want to choose p_u to maximize*

$$h(\mathbf{u}) = -\int p_u \log p_u$$

subject to the conditions that the mean is $\boldsymbol{\mu}$ and the dispersion matrix satisfies the constraint that its trace is less than or equal to ST. The solution is to choose p_u to be Gaussian with mean $\boldsymbol{\mu}$ and covariance ST/kI, where I is the identity matrix.

Proof: From the preceding theorem we only need to consider Gaussian densities. For a Gaussian density we can write the formula

$$h(u) = \frac{k}{2} \log 2\pi e + \tfrac{1}{2} \log \mid \Gamma \mid. \tag{13}$$

Maximizing $h(u)$ is equivalent to maximizing $\mid \Gamma \mid$. Now by the geometric

mean—arithmetic mean inequality

$$| \Gamma | \leqq \left[\frac{(\text{trace } \Gamma)}{k}\right]^k \tag{14}$$

with equality if and only if $\gamma_{11} = \gamma_{22} = \cdots \gamma_{kk}$.

Now we are ready to consider the finite-dimensional version of the problem of finding the optimal power-restricted signal $s_t(\omega)$ which maximizes the mutual information between it and the output $r_t(\omega)$. By what we have shown in the previous section we can replace these processes by their Fourier coefficients when computing mutual information. More specifically let $n_t(\omega)$ be a finite-dimensional Gaussian process of dimension k. Let G denote a nonsingular operator on E^k and let $s_t(\omega)$ be a k-dimensional process that is independent of $n_t(\omega)$. Suppose that the distribution of $n_t(\omega)$ is absolutely continuous with respect to Lebesgue measure in E^k. We want to find $s_t(\omega)$ such that its distribution is absolutely continuous with respect to Lebesgue measure in E^k and $I[s_t(\omega), Gs_t(\omega) + n_t(\omega)]$ is maximized subject to

$$E\left[\int_0^T s_t^2(\omega)\, dt\right] \leqq ST.$$

Now by the theorem concerning linear transformations of random vectors stated earlier, $I[s_t(\omega), Gs_t(\omega) + n_t(\omega)] = I[s_t(\omega), s_t(\omega) + G^{-1}n_t(\omega)]$. Define $\eta_t(\omega)$ as $\eta_t(\omega) = G^{-1}n_t(\omega)$ and let

$$\eta^k = \begin{Bmatrix} \eta_1 \\ \vdots \\ \eta_k \end{Bmatrix} \quad \text{and} \quad s^k = \begin{Bmatrix} s_1 \\ \vdots \\ s_k \end{Bmatrix}$$

be coordinates of $\eta_t(\omega)$ and $s_t(\omega)$.

Then

$$\begin{aligned} I[s_t(\omega), s_t(\omega) + \eta_t(\omega)] &= \int p_{s^k+\eta^k,s^k} \log \frac{p_{s^k+\eta^k,s^k}}{p_{s^k+\eta^k}p_{s^k}} \\ &= \int p_{s^k+\eta^k,s^k} \log \frac{p_{s^k+\eta^k,s^k}}{p_{s^k}} + h(s^k + \eta^k). \end{aligned}$$

Introducing the transformation

$$\begin{Bmatrix} s^k + \eta^k \\ s^k \end{Bmatrix} \to \begin{Bmatrix} \eta^k \\ s^k \end{Bmatrix}$$

into the above integral and using the fact that s^k and η^k are independent

we have

$$I(s^k, s^k + \eta^k)$$

$$= \int p_{s^k+\eta^k} \log \frac{p_{\eta^k,s^k}}{p_{s^k}} + h(s^k + \eta^k) = -h(\eta^k) + h(s^k + \eta^k). \quad (15)$$

Since $h[\eta_t(\omega)]$ is not a function of $s_t(\omega)$, we have reduced the problem to that of maximizing $h[s_t(\omega) + G^{-1}n_t(\omega)]$ subject to

$$E \int_0^T s_t^2(\omega) \leqq ST.$$

Now $\eta_t(\omega)$ is Gaussian and we know from the corollary stated earlier that $h[s_t(\omega) + G^{-1}n_t(\omega)]$ subject to the above constraint is maximized by a Gaussian process. Thus, without loss of generality, $s_t(\omega)$ can be assumed to be Gaussian. We seek Γ_s the covariance of $s_t(\omega)$ so that $|\Gamma_s + G^{-1}\Gamma_n(G^{-1})'|$ is maximized, since this maximizes $h[s_t(\omega) + G^{-1}n_t(\omega)]$. Let us assume, without loss of generality, that $G^{-1}\Gamma_n G^{-1\prime} = \Gamma_\eta$ is diagonal. Thus the problem is to maximize

$$\left| \Gamma_s + \begin{Bmatrix} \eta_1 & & 0 \\ & \ddots & \\ 0 & & \eta_k \end{Bmatrix} \right|$$

subject to Γ_s a covariance matrix with trace $(\Gamma_s) \leqq ST$. Since we are maximizing a continuous function over a compact set, we know that the maximum exists.

We use induction to show that the optimal Γ_s is diagonal. For $m = 1$ the statement is a trivial one. For $m > 1$ it shall be convenient to partition Γ_s so that

$$\Gamma_s = \begin{Bmatrix} \gamma_{11} & \gamma' \\ \gamma & \hat{\Gamma}_s \end{Bmatrix}.$$

Let

$$\Gamma = \hat{\Gamma}_s + \begin{Bmatrix} \eta_2 & & 0 \\ & \ddots & \\ 0 & & \eta_k \end{Bmatrix}.$$

Now using some standard results on determinants (see Ref. 8, p. 46), it follows that

$$|\Gamma_s + \Gamma_\eta| = \begin{vmatrix} \gamma_{11} + \eta_1 & \gamma' \\ \gamma & \Gamma \end{vmatrix} = (\gamma_{11} + \eta_1) \det \Gamma - (\det \Gamma)\gamma'\Gamma^{-1}\gamma. \tag{16}$$

Note that both Γ and Γ^{-1} are positive definite. It is optimal to choose $\gamma = 0$ to maximize the second term. The first term is also optimal if Γ is diagonal. This follows since any nondiagonal Γ with trace $\sum_{i=1}^{k} \gamma_{ii} = ST - \gamma_{11}$ has determinant less than or equal to some diagonal matrix with trace equal to $ST - \gamma_{11}$ by induction. We now optimally select the diagonal elements of Γ_s. If an $\epsilon > 0$ of ST is to be put on a diagonal element of Γ_s, it is optimal to add it to min $\{(\gamma_{ii} + \eta_i), i = 1, \cdots, k\}$ so that it will have the largest possible multiplier in the determinant of Γ.

V. CAPACITY FOR THE INFINITE-DIMENSIONAL CASE

We turn to calculating the capacity in the situation where there can he an infinite number of Fourier coefficients of $s_t(\omega)$ and $\eta_t(\omega)$ and where the channel G is an infinite-dimensional Hilhert–Schmidt operator.

Define

$$G = \int_0^T \mathbf{G}(t, \tau)$$

where $\mathbf{G}(t, \tau) \in L_2(t \times \tau)$. Let $\{\phi_i\}$ he a complete set of orthonormal eigenfunctions for $G * G$ and let $\{\lambda_i\}$ he the associated eigenvalues. Define $G\phi_i = \psi_i$; then $(\psi_i, \psi_j) = (G\phi_i, G\phi_j) = (\phi_i, G * G\phi_j) = \lambda_i \delta_{ij}$ and so $\{\psi_i/\lambda_i^{\frac{1}{2}}\}$ is an orthonormal set. We use r and η to denote the infinite vector of Fourier coefficients of $r_t(\omega)$ and $\eta_t(\omega)$ in the system $\{\psi_i/\lambda_i^{\frac{1}{2}}\}$ while $\hat{s}$ denotes the infinite vector of Fourier coefficients of $s_t(\omega)$ in the system $\{\phi_i\}$. Let r^k denote the first k coefficients of r and define $\hat{s}^k$ and η^k similarly. Let D be the douhly infinite diagonal matrix with $\lambda_i^{\frac{1}{2}}$ as the ith diagonal element and define D_k to be the $k \times k$ submatrix of D with indices less than or equal to k. Then $r = D\hat{s} + \eta$ and $r^k = D_k\hat{s}^k + \eta^k$.

We first show that if an optimal input signal exists, then there is an optimal Gaussian input signal. We shall need the following lemma.

Lemma: *For any signal* $\hat{s}$, $\lim I(r^i, \hat{s}^i) = I(r, \hat{s})$.

Proof: We know that $\lim_i \lim_k I(r^i, \hat{s}^k) = I(r, \hat{s})$. As stated earlier this is proved in Ref. 4. Now

$$I(r^i, \hat{s}^i) = h(r^i) + \int p_{r^i, \hat{s}^i} \log \frac{p_{r^i\hat{s}^i}}{p_{\hat{s}^i}}$$

where $r^i = D_i \hat{s}^i + \eta^i$. If $j \geqq i$,

$$\int p_{r^i \hat{s}^j} \log \frac{p_{r^i \hat{s}^j}}{p_{\hat{s}^j}} = \int p_{\eta^i} p_{\hat{s}^j} \log p_{\eta^i} = \int p_{\eta^i} \log p_{\eta^i} .$$

Thus $I(r^i, \hat{s}^j) = I(r^i, \hat{s}^i)$ for $j > i$; then $\lim_{j \to \infty} I(r^i, \hat{s}^j) = I(r^i, \hat{s}^i) = I(r^i, \hat{s})$. Finally $\lim_{i \to \infty} I(r^i, \hat{s}) = \lim_{i \to \infty} I(r^i, \hat{s}^i) = I(r, \hat{s})$.

Alternately this lemma can be proved by extending some results of Ref. 3 to the infinite-dimensional case.

We now show that if an optimal signal exists for the infinite-dimensional case, then the optimal signal can be assumed, without loss of generality, to be Gaussian.

Proposition III: *If $\hat{s}_1$ is a non-Gaussian optimal signal then $\hat{s}_2$, the Gaussian signal with the same covariance matrix as $\hat{s}_1$, is optimal.*

Proof: Clearly $I(r_1^k , \hat{s}_1^k) \leqq I(r_2^k , \hat{s}_2^k)$ for all k since the Gaussian process is optimal for a fixed covariance matrix in the finite-dimensional case. Thus $I(r_1 , \hat{s}_1) = \lim_k I(r_1^k , \hat{s}_1^k) \leqq I(r_2 , \hat{s}_2) = \lim_k I(r_2^k , \hat{s}_2^k)$.

Proposition IV: *The capacity of the infinite-dimensional channel is the limit of the capacities of the k dimensional truncated approximation of the infinite-dimensional channel.*

Proof: Let C_k denote the capacity of the k dimensional channel and let $C = \lim C_k$. We claim the capacity of the infinite-dimensional channel is C. It is evidently at least C. Suppose a signal $s_t(\omega)$ exists satisfying the constraints with mutual information, $I(r, \hat{s})$ greater than C. Then $I(r^k, \hat{s}^k) \leqq C_k$ since $\hat{s}^k$ satisfies the power constraint. Thus $I(r, \hat{s}) \leqq C$, a contradiction.

Corollary 1: *There exist finite-dimensional signals whose resulting mutual information is arbitrarily close to the capacity.*

Corollary 2: *If C_k is constant for all k larger than some integer l, then the l + 1-dimensional optimal signal is optimal for the infinite-dimensional case.*

5.1 *Limiting Covariance Matrices and Optimal Signals When $\{\eta_i\}$ Is Independent*

It is not always true that some input signal achieves capacity in the infinite-dimensional case. We first prove this. Then we study the special case when $\{\eta_i\}$ is independent in the $\{\psi_i/\lambda_i^{\frac{1}{2}}\}$ system. This case may be of marginal interest insofar as a model of a realistic system. However

it is mathematically tractable and hence serves as a good testing ground for intuition into more general behavior.

We now show that no optimal input signal exists for the case $\lambda_i = 1/i^2$, $E\eta_i^2/\lambda_i = 1$. It is clear that $C_k = \frac{1}{2} \log (1 + ST/k)^k$. Then the capacity is: $C = \frac{1}{2} \lim_{k\to\infty} \log (1 + ST/k)^k = ST/2$. If there exists an optimal signal s, $I(r^i, \hat{s}^i) \to ST/2$. But $I(r^i, \hat{s}^i) = \frac{1}{2} \log | \Gamma_{\hat{s}^i} + I_i |$, where I_i is an $i \times i$ identity matrix. Then $I(r^i, \hat{s}^i) \leqq \frac{1}{2} \sum_{j=1}^{i} \log (1+Es_j^2)$ and $\lim I(r^i, \hat{s}^i) \leqq \frac{1}{2} \sum_{j=1}^{\infty} \log (1 + Es_j^2)$. Recall that $\sum_{j=1}^{\infty} Es_j^2 = ST$ by assumption. We show that $\sum_{j=1}^{\infty} \log (1 + Es_j^2) < ST$. Since $Es_j^2 \geqq \log (1 + Es_j^2)$ with equality if and only if $Es_j^2 = 0$, $\lim I(r^i, \hat{s}^i) \leqq \frac{1}{2} \sum_{j=1}^{\infty} \log (1 + Es_j^2) < \frac{1}{2} \sum_{j=1}^{\infty} Es_j^2 = ST/2$.

Although an optimal signal does not always exist, we can say when it does exist in the special case when the $\{\eta_i\}$ are independent in the system $\{\psi_i/\lambda_i^{\frac{1}{2}}\}$. It will turn out that $\{\hat{\Gamma}_{\hat{s}^k}\}$, the sequence of finite-dimensional optimal covariance matrices for s, converges in some cases to an optimal solution and in other cases the limit is not optimal. The diagonal matrix with $a_i = (1/\lambda_i)E\eta_i^2$ on the ith diagonal element completely determines whether or not an optimal limit is reached.

We define the order of minima of a sequence $\{\xi_i\}_{i=1}^{\infty}$ as follows. The order is 0.5 if no smallest element in $\{\xi_i\}_{i=1}^{\infty}$ exists. If M_1 is defined to be the set of smallest elements in $\{\xi_i\}_{i=1}^{\infty}$ and $Card\ (M_1) = +\infty$, the order of the sequence is 1. If $Card\ (M_1) < +\infty$ but the set $\{\bigcup \xi_i - M_1\}$ has no least element, the order is 1.5. If the set $\{\bigcup \xi_i - M_1\}$ has M_2 smallest elements and $Card\ (M_2) = +\infty$, the order is 2. If $Card\ (M_2) < +\infty$ but the set

$$\left\{\bigcup \xi_i - \bigcup_{j=1}^{2} M_j\right\}$$

has no least element then the order is 2.5, and so on. If the sequence is not assigned a finite order of minima, the order is infinite.

If the order of minima of $\{a_i\}$ is 0.5, $\{\hat{\Gamma}_{\hat{s}^k}\} \to [0]$. To see this we need only consider diagonal elements of $\hat{\Gamma}_{\hat{s}^k}$. Suppose for some j and for some $\epsilon > 0$, $E\hat{s}_j^2 \geqq \epsilon$ in an infinite number of $\hat{\Gamma}_{\hat{s}^k}$. Since no smallest element in $\{a_i\}$ exists, there are an infinite number of a_i, say $\{a_{i'}\}$ smaller than a_j. Then in the optimal covariance matrices where $E\hat{s}_j^2 \geqq \epsilon$ and the i' appear, $Es_{i'}^2 \geqq \epsilon$. But this is not possible with the constraint $\sum E\hat{s}_i^2 \leqq ST$. Thus for each j and $\epsilon > 0$, $E\hat{s}_j^2 < \epsilon$ in all but a finite number of $\hat{\Gamma}_{\hat{s}^k}$.

If the order of the minima of $\{a_i\}$ is 1, $\hat{\Gamma}_{\hat{s}^k} \to [0]$. This follows since it is optimal to put the power on the minima. After some k, only the minima will have positive $E\hat{s}_i^2$. Since there are an infinite number of

them and the ST is optimally distributed equally on them, $\hat{\Gamma}_{\mathfrak{z}k} \to [0]$.

If the order of the minima of $\{a_i\}$ is 1.5, there are two cases. Let $h = \inf \{\bigcup \xi_i - M_1]$ and $g = \inf \{\bigcup \xi_i\}$. If $(h - g)$ $Card$ $(M_1) \geqq ST$, there is an optimal solution as a limit consisting of $E\hat{s}_i^2 = ST/Card\ (M_1)$ for those i corresponding to $a_i \varepsilon M_1$. Otherwise the covergence is to a matrix where $Es_i^2 = h - g$ if $a_i \varepsilon M_1$ and zero elsewhere, which is clearly not optimal. The analysis for other finite order systems are analogous to the above. Either (*i*) ST is distributed over a finite number of components, in which case the convergence is to an optimal solution, or (*ii*) ST has to be distributed over an infinite number of components, in which case the convergence is not to an optimal solution.

If the order is infinite and we run out of the quantity ST on a finite number of components, the resulting finite-dimensional solution is optimal. Suppose the order is infinite and we do not run out of ST on a finite number of dimensions. Let θ be the smallest accumulation point of $\{a_i\}$. If not all of ST is used in making

$$\frac{Er_i^2}{\lambda_i} = \theta,$$

the limiting covariance is not optimal and no optimal covariance which achieves capacity may be constructed. This follows since a finite amount of ST must be distributed equally to an infinite number of components. If all of the ST is exactly used to make

$$\frac{Er_i^2}{\lambda_i} = \theta,$$

the limiting covariance is optimal. Before proving this we give an example of such a case.

Let

$$\lambda_i = \frac{(i+1)}{(i+1)^3 - 1}, \qquad E\eta_i^2 = \frac{1}{(i+1)^2}$$

and assume the η_i are independent. Then $a_i = 1 - 1/(i+1)^3$. To bring all components

$$\frac{Er_i^2}{\lambda_i}$$

to 1 we need

$$ST = \sum_{i=1}^{\infty} \frac{1}{(i+1)^3},$$

and we are then in the case considered above.

We now show that the limiting covariance matrix for the case when there is just enough ST to bring

$$\frac{Er_i^2}{\lambda_i} = \theta$$

is optimal. Let $\Gamma_{\hat{s}_1}$ be the limiting matrix with the corresponding Gaussian process $\hat{s}_1$. Let $r_1 = D\hat{s}_1 + \eta$. We show that $I(r_1^k, \hat{s}_1^k) \to C$, $k \to \infty$. Now suppose $\hat{s}^k$ is optimal for k-dimensions. Then

$$I(r^k, \hat{s}^k) - I(r_1^k, \hat{s}_1^k) = h(r^k) - h(r_1^k) = \tfrac{1}{2} \log \left[\theta + \frac{\delta(k)}{k}\right]^k \prod_{i=1}^{k} \lambda_i - \tfrac{1}{2} \log \theta^k \prod_{i=1}^{k} \lambda_i . \tag{17}$$

Here $\delta(k)$ equals that part of ST not used in the matrix Γ_{s_1} in the first k-dimensions. Clearly we are assuming that the smallest elements of a_i appear first. Notice that $\delta(k) \to 0$ as $k \to \infty$. Then

$$I(r^k, \hat{s}^k) - I(r_1^k, \hat{s}_1^k) = \frac{k}{2} \log \left[1 + \frac{\delta(k)}{k\theta}\right] \leqq \frac{\delta(k)}{2\theta} \tag{18}$$

for k sufficiently large. Then

$$\lim I(r^k, \hat{s}^k) = \lim I(r_1^k, \hat{s}_1^k) = I(r_1, \hat{s}_1) = C. \tag{19}$$

VI. SUMMARY

Let us review what we have done. Since we chose to deal with signals $\xi_t(\omega)$ square-integrable on $L_2\{(\Omega, S_\Omega, P) \times ([0, T], L, m)\}$, we define the mutual information between two such signals using Proposition II and equation (2) in such a way that it agrees with the mutual information of their Fourier coefficients defined in equation (10). For the channel defined in equation (1) with input signals constrained by equation (11), we calculate the capacity of the channel. First in Section IV the capacity problem is considered when only a finite number of Fourier coefficients are nonzero. We use the corollary to the theorem in Section IV and equation (15) to show that only Gaussian signals have to be considered. Then equation (16) is used to calculate the finite-dimensional optimal signal by "filling the well." In Section V the case of an infinite number of nonzero Fourier coefficients is considered. We show in Proposition III that optimal signals, if they exist, can be chosen Gaussian. In Proposition IV the capacity of the infinite-dimensional channel is calculated as the limit of finite-dimensional capacities.

Finally in Section 5.1 we deal with the existence of an optimal signal. In general no optimal signal exists. A special case is examined when the noise components are independent in a fixed coordinate system.

APPENDIX

Symbols Used

The following is a list of symbols used throughout the text.

L_2—the set of square-integrable functions
$n_t(\omega)$—a noise process
$\eta_t(\omega)$—a noise process
$s_t(\omega)$—the input signal process
$r_t(\omega)$—the output process
G—the linear channel operator
G^*—the adjoint of G
P_ξ—a probability measure generated by ξ
p_η—the probability density of the random vector η
$\mathfrak{A}$—a sigma field
$I(\xi, \eta)$—the mutual information between ξ and η
$h(\eta)$—the differential entropy of η
Γ_s—the covariance of s
E^k—Euclidean k-space
$|\Gamma|$—the determinant of Γ
L—the Lebesgue measurable sets
m—Lebesgue measure
Card—Cardinality
E—expected value

REFERENCES

1. Lloyd, S. P., "On a Measure of Stochastic Dependence," Теория вероятностей и ее применения (Theory of Probability and Its Application), 7 (1962), pp. 312–322
2. Ghurye, S. G., "Information and Sufficient Subfields," Annals of Math. Stat., *39*, No. 6 (December 1968), pp. 2056–2066.
3. Gelfand, I. M., and Yaglom, A. M., "Calculation of the Amount of Information About a Random Function Contained in Another Such Function," Translations of the Amer. Math. Soc., (June 1959), pp. 199–246.
4. Pinsker, M. S., *Information and Information Stability of Random Processes*, San Francisco: Holden Day, 1964.
5. Doob, J. L., *Stochastic Processes*, New York: John Wiley, 1953.
6. Natanson, I. P., *Theory of Functions of a Real Variable*, New York: Ungar, 1955.
7. Rao, C. R., *Linear Statistical Inference and Its Applications*, New York: John Wiley, 1965.
8. Gantmacher, F. R., *The Theory of Matrices*, New York: Chelsea, 1959.